Spring 2026. BE_CHE_481_581 (100 points)

Homework 2. *Please submit files as an archive. This is not strictly required, but it is good practice, and it is appreciated.*

Problem 1. (50 points). To optimize the allocation of government resources, you are tasked with using the provided training dataset to train a linear regression model to predict hospital costs associated with a specific infectious disease. For predictions, use a separate inference dataset file.

Instructions:

1. **Build a single Python script** (this is not a Python-learning class—you are encouraged to seek help from AI tools) to train and predict using a linear regression model. Utilize the two provided dataset files: one labeled dataset (for splitting) and one unlabeled inference dataset. Use a 70/15/15 train/validation/test split. Report MSE, RMSE, and R^2 metrics for training, validation and test set. What would be the signs of overfitting, good generalization and underfitting? Answer in 3–5 sentences, referencing your computed metrics.

   After validation, retrain on full labeled data for final inference (no metrics needed).

   For predictions, make sure that the hospital expenses are not negative (if so, then =0). Predict the average, maximum, and standard deviation of hospital costs for the inference set of patient data.

4. Inspect regression coefficients and discuss correlations.
5. Tabulate the results in **Excel**.
6. Report the results by providing:
   - The raw inference data as an output CSV file,
   - Tables of determined values, and
   - A discussion paragraph analyzing the results.

For code and data, see attached files as “Homework_2_Q1.py” and “Homework2_Problem1_compiled_data.xlsx”

The linear regression model demonstrated strong and consistent performance across the training, validation, and test sets. The R^2 values were approximately 0.75 for the training set (0.7486), 0.74 for the validation set (0.7377), and 0.75 for the test set (0.7477), indicating that the model explains roughly 74–75% of the variance in hospital costs. The similarity in MSE and RMSE values across all three datasets suggests good

generalization, as there is no significant increase in validation or test error compared to training error. Therefore, the model does not exhibit signs of overfitting, which would be indicated by substantially lower training error relative to validation and test error. Additionally, the model does not appear to underfit, as the R^2 values are relatively high and errors are reasonably low. The regression coefficients align with domain expectations, with pre-existing conditions, occupation risk, pollution exposure, and age increasing predicted hospital costs, while vaccination status and access to healthcare significantly reduce costs.

Problem 2 (50 points): A chemical manufacturing process is controlled by two temperatures (columns 1 and 2) in °C and pressure (column 3) in kPa. The target variable y is the "Quality Rating," representing the overall quality of the product (the higher the value, the better the quality). Explore polynomial regression hypotheses, starting from linear regression and progressing to higher-order polynomials. Determine the optimal polynomial order that reasonably prevents overfitting (you can stop testing higher orders once overfitting becomes apparent). Ideally, you should loop through several polynomial orders in the same script to evaluate their impact on model performance. The features can interact, causing the appearance of cross terms in the regression (e.g. $x_1 x_2, x_2 x_3, x_1 x_3$ in the second order polynomial regression). Report MSE and R^2 for training and test sets. Plot MSE and R^2 (use matplotlib.pyplot) to visually inspect for overfitting.

0. Objective

The goal of this analysis was to model the relationship between manufacturing input variables and product **Quality** using:

- Plain Polynomial Regression (degrees 1 – 8)
- Ridge Regression (L2 regularization)
- Lasso Regression (L1 regularization)

Model complexity was varied using polynomial degree, and performance was evaluated using:

- Mean Squared Error (MSE)
- Coefficient of Determination (R^2)

Cross-validation for regularization models was performed only on the training set, and final performance was evaluated on a held-out test set.

Analysis was done utilizing the graphs, raw data can be found a\in “Homework2_Problem2_compiled_data.xlsx”. Code is attached as “Homework_2_Q2.py”

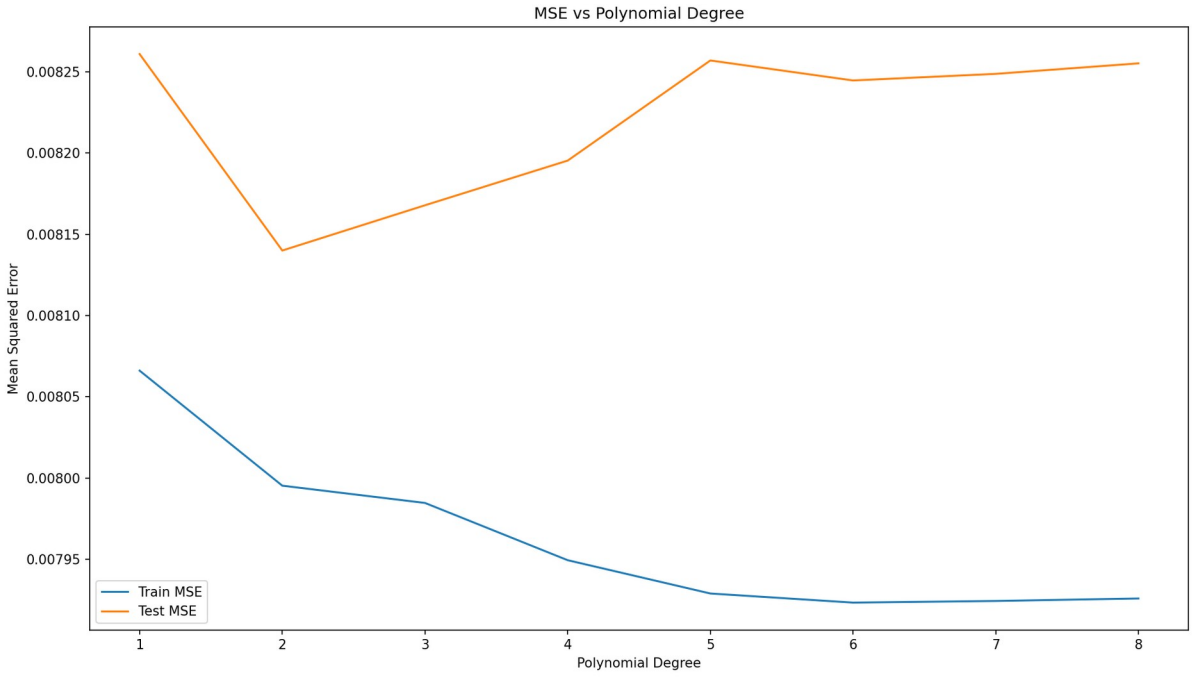
1. Plain Polynomial Regression

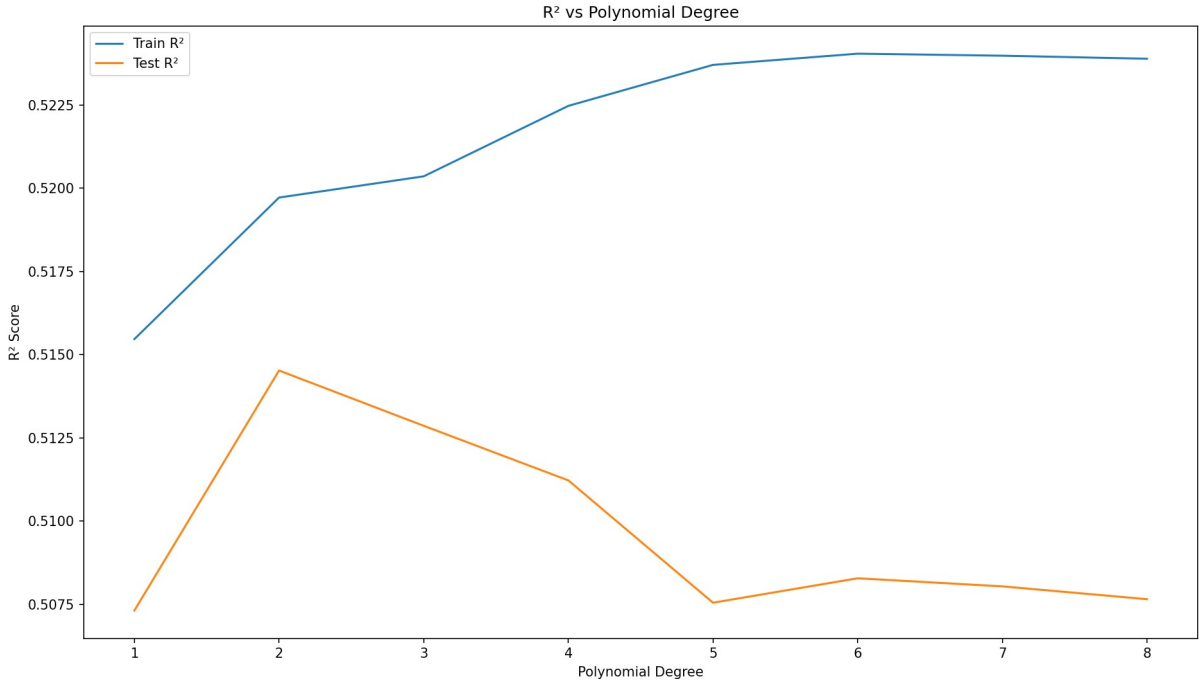
As polynomial degrees increased:

- Training MSE decreased monotonically.
- Training R^2 increased steadily toward ~0.55.
- Test MSE initially decreased, then began increasing at higher degrees.
- Test R^2 peaked at the second degree and declined afterward.

At high degrees (for this model, above degree 2), the model must have begun fitting noise in the training data, resulting in degraded test performance.

This confirms that increasing model complexity beyond a certain point harms generalization.





2. Ridge Regression (L2 Regularization)

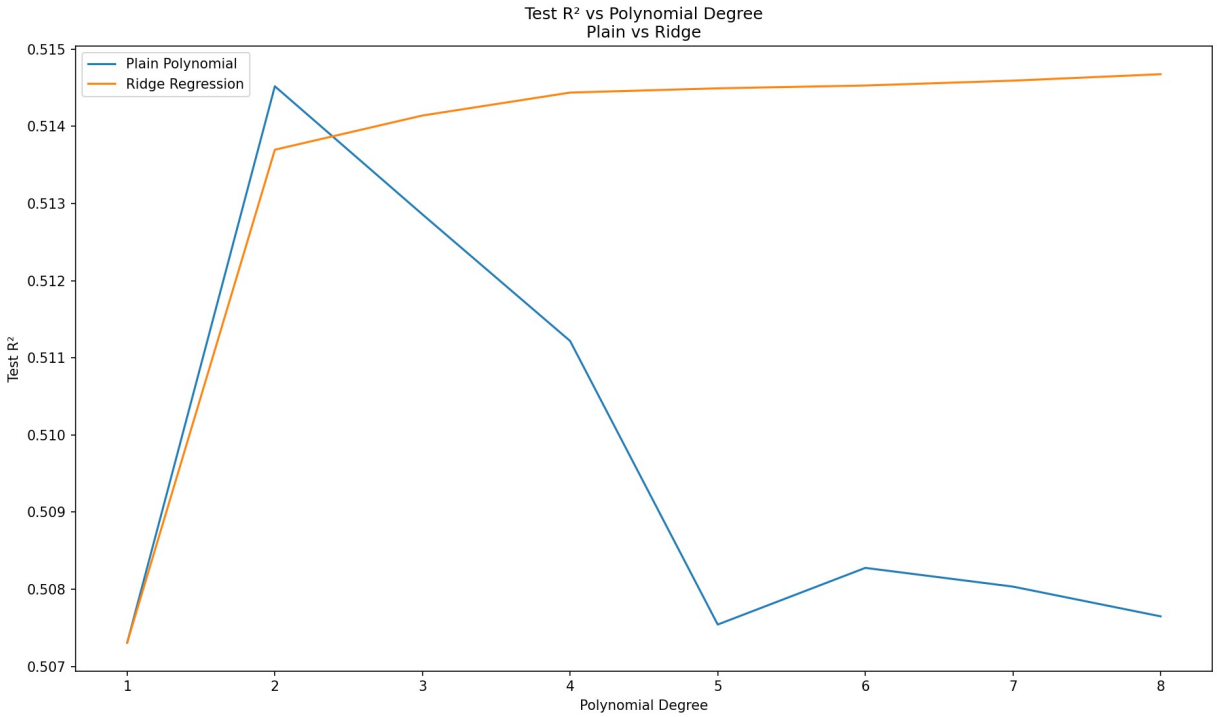
Ridge regression introduced an L2 penalty term that shrinks coefficient magnitudes but does not force them to zero.

Observations:

- Ridge reduced the performance gap between training and test error.
- Test R^2 remained more stable at higher polynomial degrees.
- Overfitting was significantly reduced compared to the plain polynomial model.

Cross-validation selected optimal alpha values that increased with polynomial degree, indicating stronger regularization was needed for more complex models.

Ridge effectively controlled variance while preserving most predictive structure.



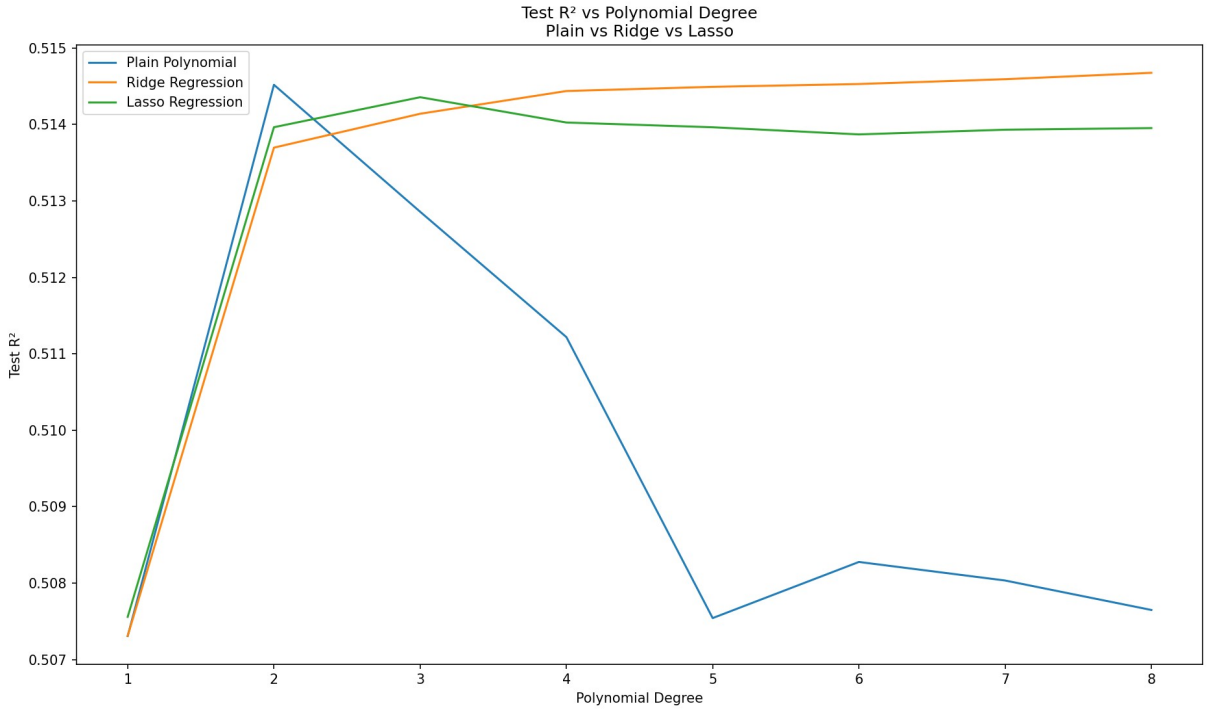
3. Lasso Regression (L1 Regularization)

Lasso regression applies an L1 penalty, which can shrink some coefficients exactly to zero.

Observations:

- Lasso performed similarly to Ridge at all degrees.
- At higher degrees, Lasso often outperformed plain polynomial models.
- The selected alpha values increased as degree increased, suggesting stronger shrinkage was required to control complexity.

Because Lasso performs feature selection, it likely eliminated unnecessary high-order interaction terms.



4. Model Comparison

Comparing the best-performing configuration of each model:

- Plain Polynomial achieved a really strong performance at degree 2 but degraded at high degrees.
- Ridge provided consistent improvement over the plain model at higher degrees, it's best R^2 at degree 8.
- Lasso offered the strongest regularization effect, peaking at degree 3 and then stabilizing afterwards.

Overall, the regularized models demonstrated superior generalization performance compared to the unregularized polynomial model, particularly at higher degrees.

5. Final Model Selection

Based on highest test R^2 and lowest test MSE:

The best-performing model was:
Ridge

This model achieved the best balance between model complexity and generalization, having the highest R^2 value at the 8^{th} degree and ensuring a low MSE at all points.